Phytoplankton Biomass Predictions in Equatorial Pacific using a Random Forest Model

by

Cristian Swift

Email: Crisjay01@gmail.com

University of Washington

School of Oceanography

February 20th, 2023

### **Project Summary**

Phytoplankton Size Class (PSC) is an important observation that can inform us of sequestration of surface carbon to the deep ocean in addition to environment nutrient load, photosynthetic efficiency, and even the marine food web structure. In the equatorial Pacific, few oceanographic cruises have been conducted, especially those that have continuous measurements of PSC using SeaFlow, leaving a gap in knowledge. Observational variables from cruise underway nitrate, satellite derived Chl a and sea surface temperature, and ENSO conditions will be used as features in a Random Forest Classifier model to predict PSC on a given cruise tract using above observational features.

### **Introduction**

Phytoplankton size class (PSC) are good proxies for sequestration of surface carbon to the deep ocean. From Stokes’ Law we observe that a particle's radius is directly proportional to the sinking rate, so smaller phytoplankton sink at a smaller rate than larger phytoplankton. This impacts the “biological pump” , the process in which atmospheric CO2 dissolved in seawater is incorporated into organic matter through photosynthesis. Once the organism goes through its life cycle, a portion of the organic matter descends from the surface oceans as marine snow (organic material falling from upper waters to the deep ocean). Marine snow sinking to depth results in atmospheric CO2 “pumped’ and sequestered at depth.

The rate at which they sink is partially dependent on the organic matter’s surface area to volume ratio as explained by Stokes Law, where a particle with a larger surface area to volume ratio sinks (smaller PSC) at a lower rate (Turner, 2015). Phytoplankton are diverse physiologically with size ranges from 0.6 to 200 µm, with three size classes, picoplankton (diameter <2 μm), nanoplankton (diameter from 2 to 20 μm), and microplankton (diameter >20 μm) (Brotas et al., 2022). While smaller PSC generally occupy the mixed layer longer due to their lower sinking rate, their carbon export is proportional to their contributions to total net primary productivity just as larger PSC. Such pathways of export for picophytoplankton include aggregation and their incorporation into detritus as marine snow. Additionally, picophytoplankton consumption through higher trophic levels leads to carbon export indirectly (Richardson and Jackson 2007). Dune et al., 2005 also demonstrated that carbon export is relative to productivity, with larger phytoplankton, such as diatoms, that utilize mineralization, being more productive than smaller cells, resulting in increased carbon export to depth.

PSC gives insight into environment nutrient load, photosynthetic efficiency, and even the marine food web structure. As nutrient load decreases in the marine environment there is a trend of decreasing PSC, smaller plankton are favored in oligotrophic oceans since they have a higher surface area to volume ratio making passive nutrient uptake proportionally higher in terms of size than larger phytoplankton (Acevedo-Trejos et al., 2018). Nitrate is a significant limiting nutrient for these organisms and its concentrations are correlated with other necessary nutrients as observed in the Redfield ratio (Tyrrell 2019). These correlations make nitrate a cool proxy for estimating nutrient concentrations in the water column. Robinson et al., 2018 found that mean maximum photosynthetic rates have a positive relationship with PSC, meaning smaller size classes plankton (pico and nano) have higher photosynthetic rates than their larger counterparts. These smaller phytoplankton also have increased photosynthetic efficiency, allowing for photosynthesise at lower light levels than larger phytoplankton. Additionally it was noted that maximum photosynthetic rates and mean light limited slope increase with temperature in micro-plankton, but there is a negative relationship with pico-plankton, and a positive relationship of mean maximum photosynthetic rates with temperature and no correlation of mean light limited slope. Meaning in waters with warmer Sea Surface Temperature (SST) micro-plankton will have higher photosynthetic rates than pico-plankton. As atmospheric CO2 increases globally, since the industrial revolution, due to anthropogenic output from the use of fossil fuels, both atmospheric and SST increase. Additionally, the El Niño and the Southern Oscillation (ENSO) are a larger driver in change of SST in time scales of one to three years, with the atmospheric process heavily impacting the equatorial Pacific ocean.

The equatorial Pacific ocean is an oligotrophic marine environment that experiences relatively high SST compared to other pelagic environments due higher photon intensity and is influenced by the El Niño and the Southern Oscillation (ENSO). ENSO is a large-scale interaction between the ocean and the atmosphere and it cycles between El Niño, “normal”, and La Niña. El Niño is characterized by weakened equatorial easterlies winds, this results in negative sea level pressure and reduced upwelling. This results in increased atmospheric temperature along with higher SST due to a lack of cold, nutrient rich water upwelling.(Wang & Fiedler, 2006). ENSO is most concentrated near the equator in the Pacific, with the largest changes in SST occurring there, and its variability has been increasing since the industrial revolution (Figure 1). Due to increased temperatures, and a lack of upwelling nutrients, such as nitrate, higher populations of pico-plankton and nano-plankton are observed, such as in the El Niño event of 1997/98 where pico-plankton and nano-plankton were more populous, while microplankton made up the majority once the El Niño event had ended (Irarte & González, 2004).

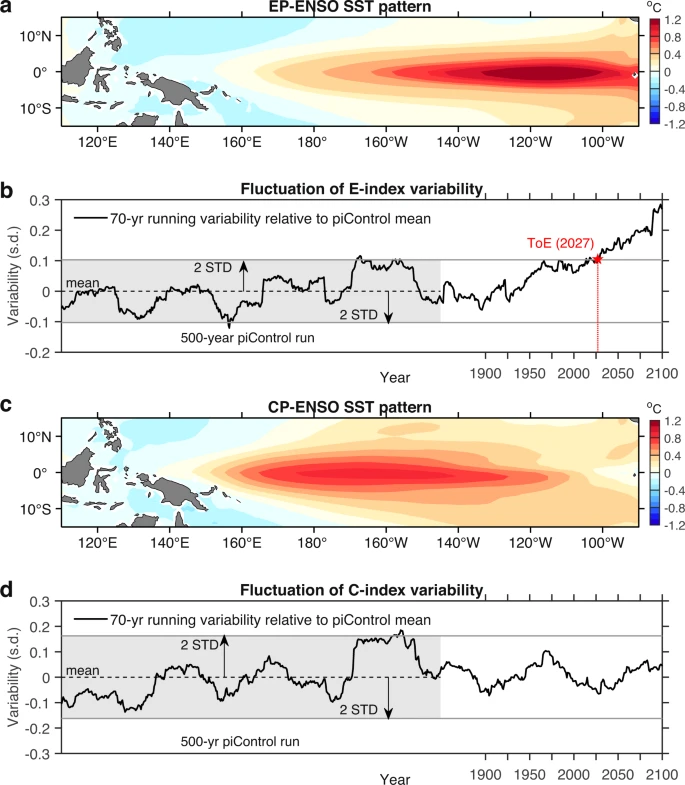


Figure 1: (Geng et al., 2022) El Niño SST variability from climate model GFDL-ESM4. Panel a shows the spatial pattern of ENSO in the eastern equatorial Pacific. Panel B describes El Niño variability relative to the mean of the E-index variability from pre-industrial levels. Panel C and D describe the same as panels A and B but relative to the central Pacific.

Few oceanographic research cruises take place in the equatorial Pacific, leaving a large gap of knowledge in PSC in the region. Especially considering global climate change leading to higher sea SST globally and increasingly frequent and intense El Niño events, it is important to understand and predict how PSC is changing and impacted by ENSO events. Satellite data has shown to be a promising remote sensing tool to predict PSC. Hu et al. 2018 found the Random Forest (RF) model to be the best retrieval model of four tested for predicting PSC based on satellite of near surface Chl a and photosynthetically active radiation (PAR). Though the models developed in Hu et al. 2018 were not specific to the equatorial Pacific. Additionally these models did not include SST and in situ measurements such as nitrate concentration in the seawater.

Through predicting PSC, we could expand PSC measurements on previous research cruises that did not have the instruments to collect this data. PSC is an important observation that can inform us of sequestration of surface carbon to the deep ocean in addition to environment nutrient load, photosynthetic efficiency, and even the marine food web structure. By expanding our knowledge of PSC through previous cruises taking place in the equatorial Pacific, we can gain a better understanding of inferred phytoplankton community changes over time.

**Proposed research:**

Hypothesis:

To address the lack of equatorial Pacific ML models, I will use nitrate, sea surface temperature, and satellite derived products (CHl a, PAR) to create an RF model to predict PSC using satellite and cruise data. Additionally ENSO conditions (El niño, “normal”, and El niña) will be incorporated into the model to determine if El niño conditions can aid in the prediction of the Random Forest model. I hypothesize that nitrate concentration and Sea surface temperature sets phytoplankton size class because PSC often changes in correlation with those parameters, additionally satellite chlorophyll data and ENSO will aid in informing PSC.

**Materials and methods:**

Data Retrieval:

SeaFlow testing and training data will be obtained using Simons CMAP with cruise data being limited to the equatorial Pacific ocean between latitudes -38.0082°, 59.9944°. [SeaFlow data](https://simonscmap.com/catalog/datasets/all_SeaFlow_cruises_v1_5) from Simons CMAP, this dataset includes 64 oceanographic cruises from 2010-05-04 to 2021-12-30, but data will be limited to the equatorial Pacific. Satellite chlorophyll color data comes from [NASA's OceanColor Web](https://oceancolor.gsfc.nasa.gov/), which will be restricted to the same tempo-spatial boundaries as the SeaFlow data. The near surface Chl a comes in 8-day composite products with a resolution of 9 x 9 km2. Sea surface temperature will also be collected from [NASA EarthData SeaBASS](https://seabass.gsfc.nasa.gov/wiki/sst_validation_description). Underway in-situ nitrate data will come from the same cruises where SeaFlow data is available, cruises will be limited to those that have underway nitrate data. SeaFlow data will be collected from the whole cruise tract and collection will be conducted by a team aboard ship.

Random Forest Model:

Random forest models work using a system of classification and regression trees, this model is trained using bootstraps which is a resampling technique that repeatedly takes samples from testing data with replacement (Sarica et al., 2017). I will use the RF classifier model since we are predicting PSC, which is not continuous, but three discrete classes. Using anaconda python software, the RF will be a classifier model from [sklearn.ensemble.RandomForestClassifier](http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html). 70 percent of the data will be used for training, and 30 percent of the data will be tested on by the resulting model. The label will be PSC and the features will be nitrate conc., SST, Chl a, ENSO conditions. Feature selection will be incorporated to develop a simpler and more effective model. Using the feature\_importances\_attribute of the RandomForestClassifier class will rank the features in terms of their importance to aiding the model in predicting the label. This will both tell us which features influence PSC most and aid us in simplifying the model through the removal of features.

In order to assess the performance of the model, I will use a 10-fold cross validation compute the coefficient of determination (R2), root mean square error (RMSE), mean absolute percentage error (MAPE), and relative RMSE (RRMSE) using cross\_val\_score from sklearn.model\_selection.

Begin developing Random Forest model on data obtained from Simon CMAP using most data for training and some of it for testing. Run statistical analysis to properly prepare data for training and testing. Process and prepare data from SST, Nutrients, and SeaFlow data into an appropriate data frame. Run statistical analysis on the data frame to understand any trends and filter any non-actual values. Then begin applying the model to nutrient, SST, and SeaFlow data to the model. Run statistical analysis on the model on the cruise data to determine models’ performance and visualize performance. Collect results and incorporate them into a research paper. Develop multiple drafts from faculty and peer reviews. Finalize paper by May 26th.

**Results**

The Random Forest model on initial testing received a mean absolute error of 4.5 pgC/L when used on the testing dataset. Additionally using sklearns accuracy rating function, it was determined the random forest model has an accuracy of -73%. Figure one demonstrates no single clear relationship between true and predicted biomass when used on the whole available dataset. However, there are two or three clusters of data present that have a clearer relationship.

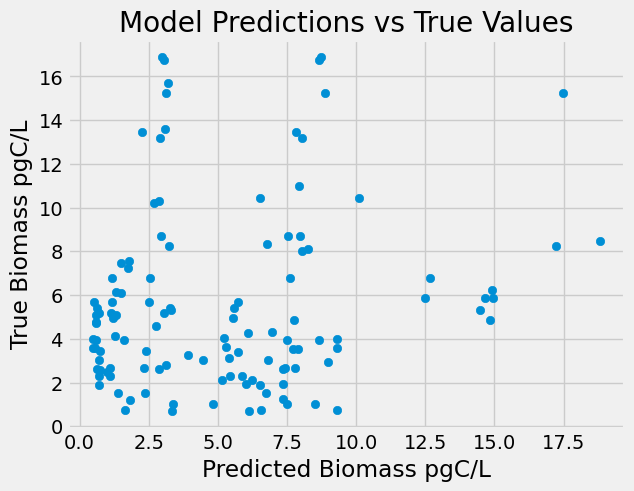


Figure : Random Forest Model True vs Predicted Biomass (pgC/L) across all populations present in the data set.

However, when the model is tested individually on the plankton populations some trends become clearer. Prochlorococcus has no clear trends when tested on true versus predicted biomass, through Synechococcus does have clearer linear relationship. The Picoeukaryotes have two different clusters of related datapoints in the lower x region and upper x region. The population that had the clearest relationship between true and predicted biomass pgC/L was the nanoeukaryotes.

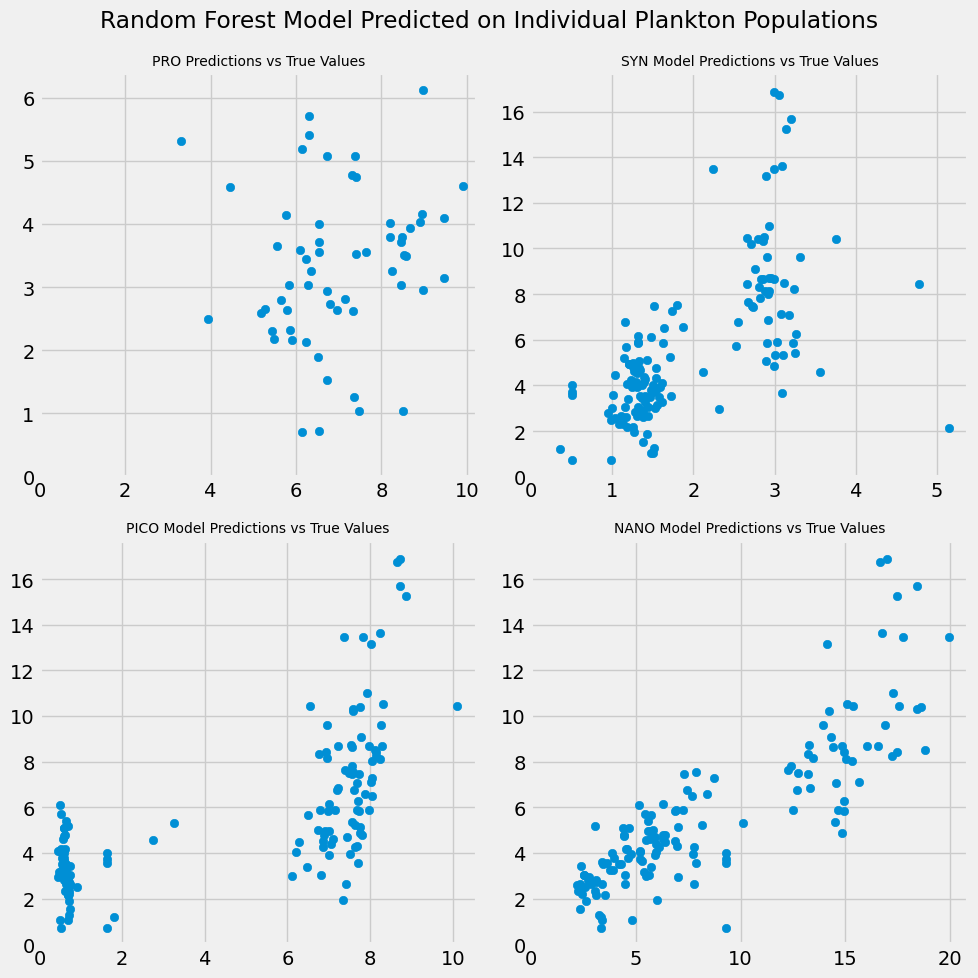


Figure : Random Forest Model Predictions on the training/testing dataset separated by phytoplankton population.

When the model takes predictions, it doesn’t use each feature equally. Figure 3 demonstrates which features the model notes as most important for deciding what the biomass of a given dataset with the correct features is. This random forest model has latitude as it largest contributor to deciding the biomass of a given set of data at 29%, following is temperature at 27%, and salinity at 21%, making those the top three most important features. The features that received the lowest importance ratings (below 5%) were all of the nutrients (besides phosphorus), along with mixed layer depth and satellite chlorophyll.

Chart

Description automatically generated

Figure : Random Forest feature importance

### **References**

### Wang, H., T. Liu, Y. Huang, Y. Fang, R. Liu, S. Wang, W. Wen, and M. Sun. 2014. Plasmon-driven surface catalysis in hybridized plasmonic gap modes. Sci. Rep. 4: 7087.

### Acevedo-Trejos, E., E. Marañón, and A. Merico. 2018. Phytoplankton size diversity and ecosystem function relationships across oceanic regions. Proceedings of the Royal Society B: Biological Sciences **285**: 20180621. doi:https://doi.org/10.1098/rspb.2018.0621

### Brotas, V., G. A. Tarran, V. Veloso, and others. 2022. Complementary Approaches to Assess Phytoplankton Groups and Size Classes on a Long Transect in the Atlantic Ocean. Frontiers in Marine Science **8**. doi:https://doi.org/10.3389/fmars.2021.682621

### Buesseler, K. O., C. H. Lamborg, P. W. Boyd, and others. 2007. Revisiting Carbon Flux Through the Ocean’s Twilight Zone. Science **316**: 567–570. doi:https://doi.org/10.1126/science.1137959

### Dunne, J. P., R. A. Armstrong, A. Gnanadesikan, and J. L. Sarmiento. 2005. Empirical and mechanistic models for the particle export ratio. Global Biogeochemical Cycles **19**: n/a-n/a. doi:https://doi.org/10.1029/2004gb002390

### Dutkiewicz, S., P. Cermeno, O. Jahn, M. J. Follows, A. E. Hickman, D. A. A. Taniguchi, and B. A. Ward. 2020. Dimensions of marine phytoplankton diversity. Biogeosciences **17**: 609–634. doi:https://doi.org/10.5194/bg-17-609-2020

### Geng, T., W. Cai, L. Wu, and others. 2022. Emergence of changing Central-Pacific and Eastern-Pacific El Niño-Southern Oscillation in a warming climate. Nature Communications **13**. doi:https://doi.org/10.1038/s41467-022-33930-5

### Hu, S., H. Liu, W. Zhao, T. Shi, Z. Hu, Q. Li, and G. Wu. 2018. Comparison of Machine Learning Techniques in Inferring Phytoplankton Size Classes. Remote Sensing **10**: 191. doi:https://doi.org/10.3390/rs10030191

### Irarte, J. L., and H. E. González. 2004. Phytoplankton size structure during and after the 1997/98 El Niño in a coastal upwelling area of the northern Humboldt Current System on JSTOR.

### Marañón, E. 2015. Cell Size as a Key Determinant of Phytoplankton Metabolism and Community Structure. Annual Review of Marine Science **7**: 241–264. doi:https://doi.org/10.1146/annurev-marine-010814-015955

### Meyer, M. G., W. Gong, S. M. Kafrissen, and others. 2022. Phytoplankton size-class contributions to new and regenerated production during the EXPORTS Northeast Pacific Ocean field deployment. Elementa: Science of the Anthropocene **10**. doi:https://doi.org/10.1525/elementa.2021.00068

### Richardson, T. L., and G. A. Jackson. 2007. Small Phytoplankton and Carbon Export from the Surface Ocean. Science.

### Robinson, A., H. A. Bouman, G. H. Tilstone, and S. Sathyendranath. 2018. High photosynthetic rates associated with pico and nanophytoplankton communities and high stratification index in the North West Atlantic. Continental Shelf Research **171**: 126–139. doi:https://doi.org/10.1016/j.csr.2018.10.010

### Sarica, A., A. Cerasa, and A. Quattrone. 2017. Random Forest Algorithm for the Classification of Neuroimaging Data in Alzheimer’s Disease: A Systematic Review. Frontiers in Aging Neuroscience **9**. doi:https://doi.org/10.3389/fnagi.2017.00329

### Srivastava, D., and Lekha Bhambhu. 2010. Data classification using support vector machine. ResearchGate.

### Swalwell, J. E., F. Ribalet, and E. V. Armbrust. 2011. SeaFlow: A novel underway flow-cytometer for continuous observations of phytoplankton in the ocean. Limnology and Oceanography: Methods **9**: 466–477. doi:https://doi.org/10.4319/lom.2011.9.466

### Swalwell, J., François Ribalet, and E. V. Armbrust. 2013. SeaFlow: A novel underway flow-cytometer for continuous observations of phytoplankton in the ocean. ResearchGate.

### Turner, J. T. 2015. Zooplankton fecal pellets, marine snow, phytodetritus and the ocean’s biological pump. Progress in Oceanography **130**: 205–248. doi:https://doi.org/10.1016/j.pocean.2014.08.005

### Tyrrell, T. 2019. Redfield Ratio. Encyclopedia of Ocean Sciences 461–472. doi:https://doi.org/10.1016/b978-0-12-409548-9.11281-3

### Wang, C., and P. C. Fiedler. 2006a. ENSO variability and the eastern tropical Pacific: A review. Progress in Oceanography **69**: 239–266. doi:https://doi.org/10.1016/j.pocean.2006.03.004

### Wang, C., and P. C. Fiedler. 2006b. ENSO variability and the eastern tropical Pacific: A review. Progress in Oceanography **69**: 239–266. doi:https://doi.org/10.1016/j.pocean.2006.03.004

### Watanabe, M., J.-L. Dufresne, Y. Kosaka, T. Mauritsen, and H. Tatebe. 2020. Enhanced warming constrained by past trends in equatorial Pacific sea surface temperature gradient. Nature Climate Change **11**: 33–37. doi:https://doi.org/10.1038/s41558-020-00933-3

### 